Lending Club Case Study

Submitted by:
Adityanarayana Rao Yerabati

Index

1.Problem Statement

The objective of this analysis is to identify high-risk loan applicants who are likely to default, thereby minimizing financial losses for a consumer finance company that provides various types of loans to urban customers.

The company faces two types of risks in its loan approval process:

1. opportunity loss from rejecting applicants who would repay their loans
2. credit loss from approving applicants who default.

The loan data includes three possible outcomes for approved loans: "Fully Paid" (loan repaid with interest), "Current" (loan still being repaid), and "Charged-Off" (loan default, the primary source of loss).

Rejected loans do not have transaction history. The company aims to identify risky borrowers to reduce credit losses by denying loans, reducing loan amounts, or charging higher interest rates for high-risk applicants.

Through Exploratory Data Analysis (EDA), the results of this analysis will be presented in a well-commented Python file with visualizations, highlighting key findings from univariate and bivariate analysis and providing business insights.

The goal of this analysis is to assist LendingClub in reducing credit losses by addressing two critical scenarios:

- Recognizing applicants who are likely to repay their loans is essential, as they contribute to the company’s profitability through interest payments. Denying these applicants would result in missed business opportunities.
- Approving loans for applicants who are unlikely to repay, and are therefore at risk of default, could lead to significant financial setbacks for the company.

2.Data Analysis

➔ The key attribute is: *loan_status* and contains 3 distinct values:

1. *Fully Paid*:
   a. Applicant has fully paid the loan (the principal and interest rate)
2. *Charged Off*:
   a. Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has **defaulted** on the loan
3. *Current*:
   a. Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.

➔ For this case study, we will exclude *Current* values because they do not provide any meaningful insights for assessing default factors.

➔ <u>The key predictors for loan approval are:</u>

1. Customer Information:
    a. *Annual income*
    b. *Home Ownership*
    c. *Employment Length*
    d. *DTI (Debt to Income)*
    e. *Address State*
    f. *Verification Status*
    g. *Purpose*

2. Customer Loan Information:
    a. *Loan Amount*
    b. *Grade and Subgrade*
    c. *Term*
    d. *Loan Issue Date*
    e. *Interest Rate*
    f. *Installment*
    g. *Public Records*
    h. *Public Records Bankruptcy*
    i. *Funded Amount*
    j. *Funded Amount Invested*
    k. *Last Payment Amount*
    l. *Total Account*

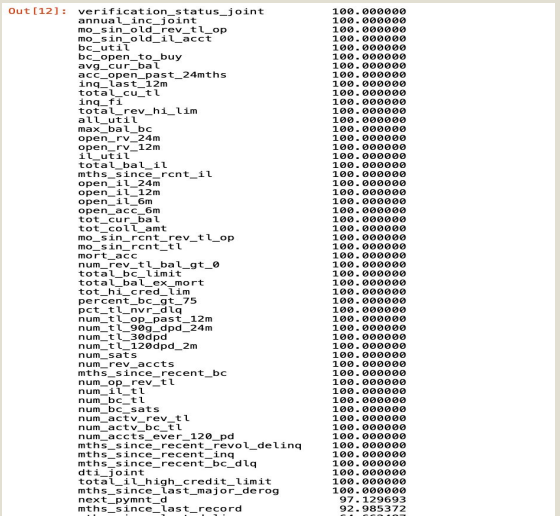
➔ Columns that will not be included in our analysis:

1. Columns with 80% or more missing data, i.e. null values are present
2. Columns consisting of NA values
3. Index columns
4. Missing data in columns
5. Columns containing only constant values which do not provide any meaningful insights to us

➔ Column Names that will be dropped:

"collection_recovery_fee", "delinq_2yrs", "desc", "emp_title", “inq_last_6mths", "last_credit_pull_d", "open_acc", "Out_prncp", “out_prncp_inv", "pub_rec", "recoveries", "title", "total_pymnt", "total_pymnt_inv", "url", "zip_code", “tax_liens, “delinq_amnt”, “chargeoff_within_12_mths”, “acc_now_delinq”, “application_type”, “policy_code”, “collections_12_mths_ex_med”, “initial_list_status”, “pymnt_plan”.

Including the columns with 80% or more missing data:



Out[12]:

verification_status_joint	100.000000
annual_inc_joint	100.000000
mo_sin_old_rev_tl_op	100.000000
mo_sin_old_il_acct	100.000000
bc_util	100.000000
bc_open_to_buy	100.000000
avg_cur_bal	100.000000
acc_open_past_24mths	100.000000
inq_last_12m	100.000000
total_cu_tl	100.000000
inq_fi	100.000000
total_rev_hi_lim	100.000000
all_util	100.000000
max_bal_bc	100.000000
open_rv_24m	100.000000
open_rv_12m	100.000000
il_util	100.000000
total_bal_il	100.000000
mths_since_rcnt_il	100.000000
open_il_24m	100.000000
open_il_12m	100.000000
open_il_6m	100.000000
open_acc_6m	100.000000
tot_cur_bal	100.000000
tot_coll_amt	100.000000
mo_sin_rcnt_rev_tl_op	100.000000
mo_sin_rcnt_tl	100.000000
mort_acc	100.000000
num_rev_tl_bal_gt_0	100.000000
total_bc_limit	100.000000
total_bal_ex_mort	100.000000
tot_hi_cred_lim	100.000000
percent_bc_gt_75	100.000000
pct_tl_nvr_dlq	100.000000
num_tl_op_past_12m	100.000000
num_tl_90g_dpd_24m	100.000000
num_tl_30dpd	100.000000
num_tl_120dpd_2m	100.000000
num_sats	100.000000
num_rev_accts	100.000000
mths_since_recent_bc	100.000000
num_op_rev_tl	100.000000
num_il_tl	100.000000
num_bc_tl	100.000000
num_bc_sats	100.000000
num_actv_rev_tl	100.000000
num_actv_bc_tl	100.000000
num_accts_ever_120_pd	100.000000
mths_since_recent_revol_delinq	100.000000
mths_since_recent_inq	100.000000
mths_since_recent_bc_dlq	100.000000
dti_joint	100.000000
total_il_high_credit_limit	100.000000
mths_since_last_major_derog	100.000000
next_pymnt_d	97.129693
mths_since_last_record	92.985372

3.Data Cleaning & Imputing

➔ The following steps are taken for data cleaning and imputing missing values:

1. Importing libraries
2. Read data
3. Gather basic information about the data
4. Removal of null values
5. Creating derived columns
6. Handling outliers
7. Imputing data

1. Importing Libraries:
   a. Declaring and importing the required libraries
2. Reading the Data:
   a. Using the correct format
3. Gathering basic information about the data:
   a. Shape
   b. Info
   c. Describe
   d. Data types

```
No of Columns: 111
No of Rows: 39717
No of missing values: 2263364
No of unique values: 416801
No of duplicates: 0
```

4. Removal of Null Values:
    a. Columns with 80% or more missing data, i.e. null values are present are dropped
    b. Rows that have single values, constant values, zero values, etc are removed from the dataset and dropped.

5. Creating Derived Columns:
    a. Columns such as “issue_d” and “earliest_cr” was split and year, month, and day was created with each being a separate column
    b. In addition, buckets were created for “loan_amnt” and “int_rate” so analysis could be performed easier.

6. Handling Outliers:
    a. The annual_inc of most of the loan applicants is between 35,000 - 80,000
    b. The loan amount of most of the loan applicants is between 5,000 - 15,000
    c. The funded amount of most of the loan applicants is between 5,000 - 13,500
    d. The funded amount by investor for most of the loan applicants is between 5,000 - 15,000
    e. The interest rate on the loan is between 9% - 14.5%
    f. The monthly installment amount on the loan is between 180 - 430
    g. The debt to income ratio is between 8 - 18

7. Imputing Missing Data:

a. “Emp_length” column was mapped to a number that it represented.
b. Imputed NONE values into OTHER in “home_ownership”
c. We took empty “emp_length” as business owners as they do not work for a company and rather don’t have an x amount of years they have worked.
d. Replaced “Source Verified” with “Verified”

The 7 steps that we have done are extremely important for data analysis because to get meaningful insights it is necessary to clean and discard values that are not important.

The final shape of the data frame is:

```
In [382]: 1 loan_df.shape
Out[382]: (37880, 35)
```

4.Univariate Analysis

Univariate analysis is the simplest form of data analysis, where a single variable is analyzed at a time to understand its distribution, central tendency (mean, median, mode), dispersion (variance, standard deviation), and range. The main goal is to summarize and describe the variable's characteristics.

Key Features of Univariate Analysis

1. **Focus:** Examines only one variable at a time.
2. **Purpose:** Helps understand the basic properties of a dataset before exploring relationships with other variables.
3. **Output:** Includes numerical measures, charts, and summaries.

Why is Univariate Analysis Important?

1. **Initial Insights:** Provides a foundation for understanding the data.
2. **Data Cleaning:** Helps detect outliers, missing values, or skewness.
3. **Guides Next Steps:** Sets the stage for bivariate or multivariate analysis.

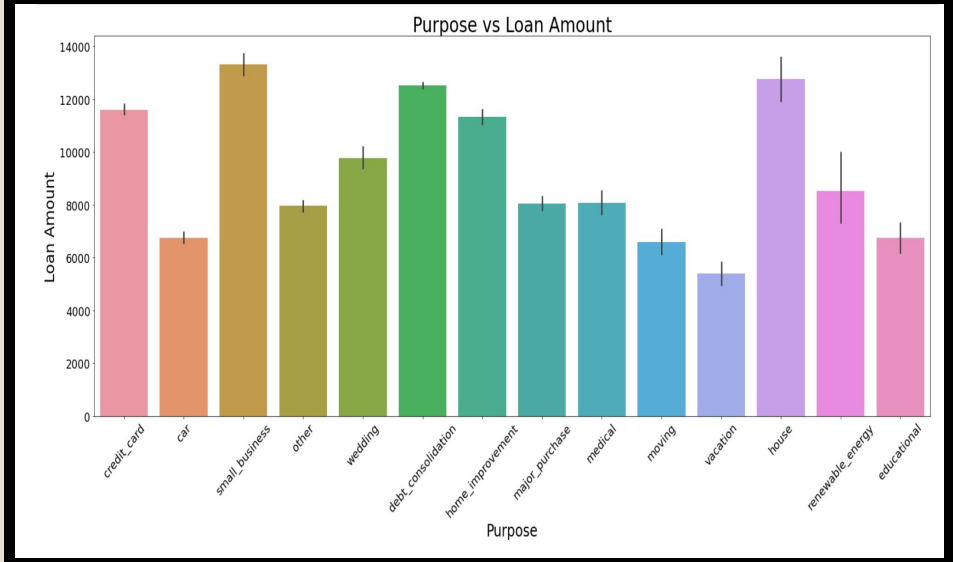
Techniques for Univariate Analysis

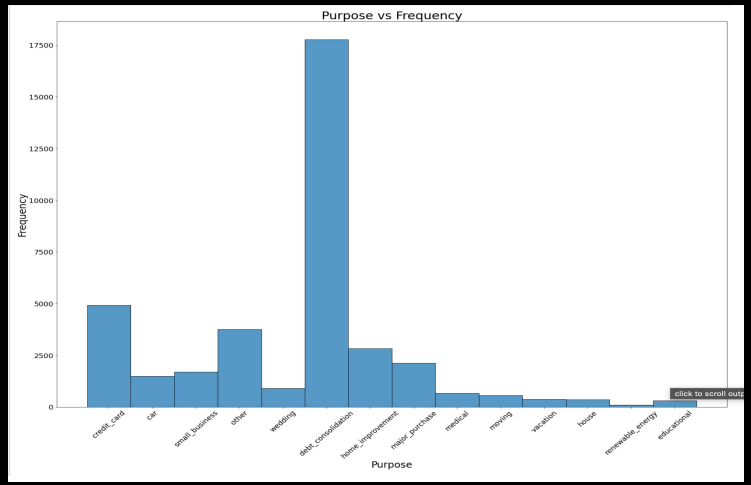
Numerical Summaries:

- **Mean:** Average value.
- **Median:** Middle value when data is sorted.
- **Mode:** Most frequently occurring value.
- **Variance and Standard Deviation:** Spread of the data.
- **Range:** Difference between the maximum and minimum values.

Inference:

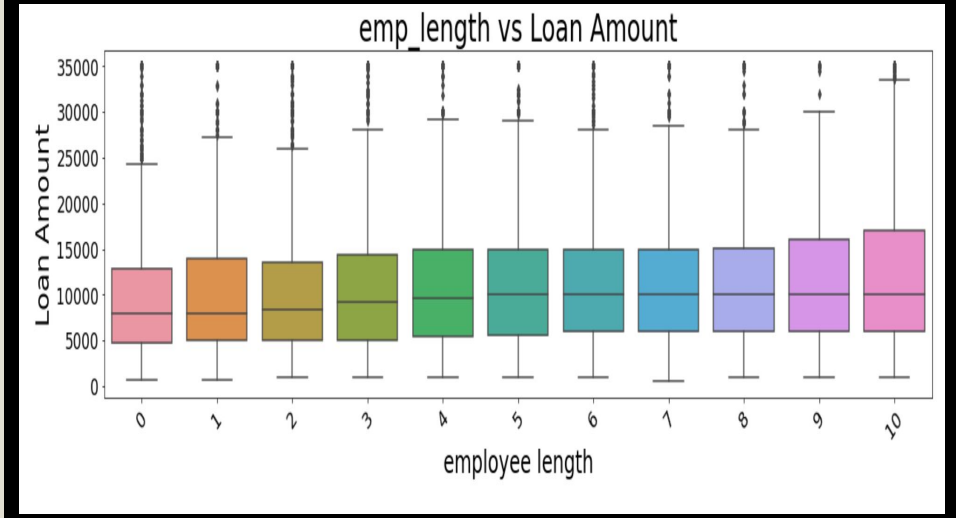
- The highest loan amount are for those who take it for the purpose of debt_consolidation, small business, and house loans.
- There are various reasons as to why a person would want to take those 3 loans, but debt_consolidation is the most popular because it involves combining multiple loans into one with a lower interest rate, thus the amount will be larger and interest rate will be less.
- As per the histogram, debt_consolidation has the highest with over 17,500.





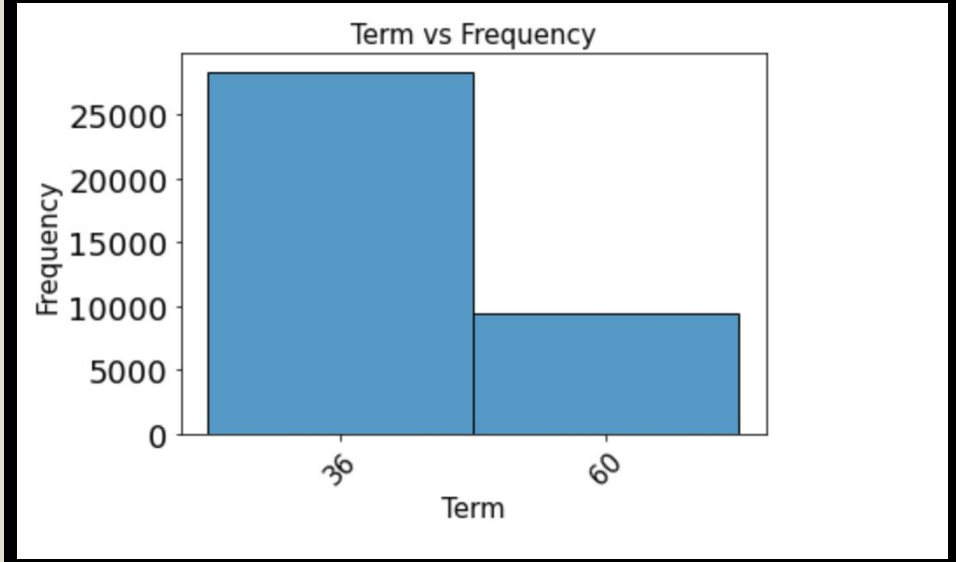
Inference:

- Based on this graph, we can analyze that employees who have an experience of 10+ years will more likely take higher loan amounts (approx 30,000-35,000 USD) compared to those who have less years of experience, i.e. below 3 years.
- And the average loan amount based on this boxplot for those who have 10 years of experience is around 10,000-20,000 USD) compared to those with less years of experience.
- Based on this, we can conclude that those who have more experience are more likely to take a higher loan amount given the experience, and salary they are compensated with compared to those who just started out with a low salary.
- One observation to note down is that the many customers take loans between 15,000-25,000 USD range.



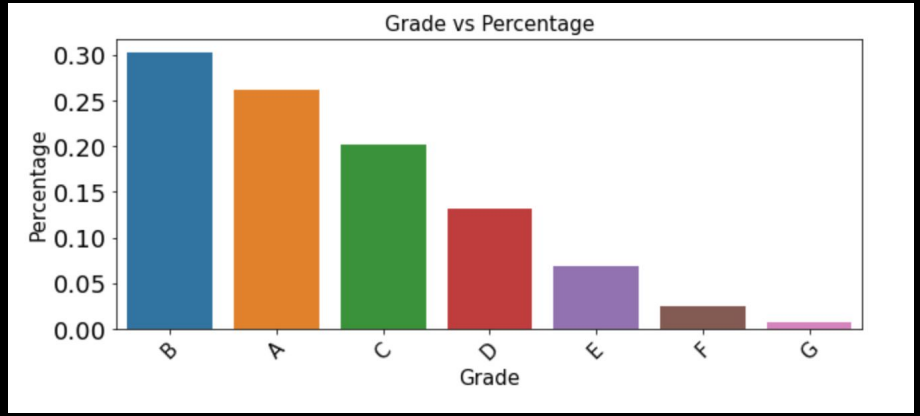
Inference:

- Based on the distribution, we can infer that the loans that customers take for 36 months are higher (25,000+) compared to the customers who take a 60 months loan (around 9000).
- And, this can be due to reasons such as less interest rate and the capability of the customer to pay it back in a shorter amount of time.



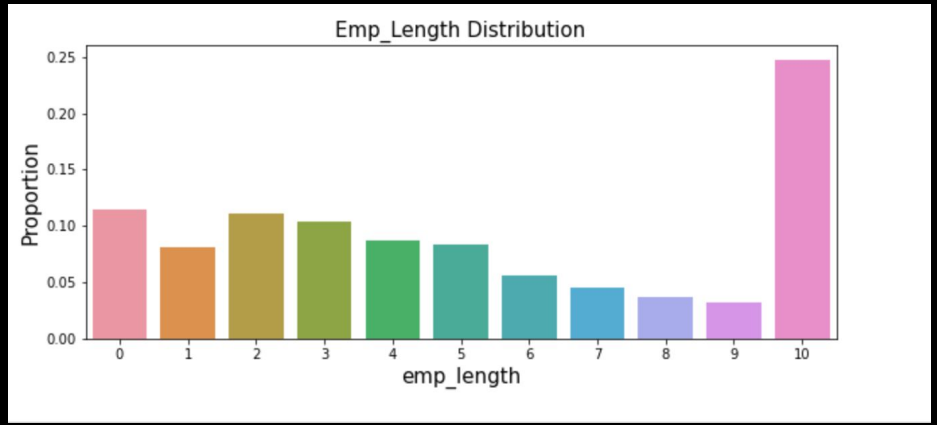
Inference:

- Based on the “Grade” vs “Percentage” graph we can infer that those who got a grade **B** had a higher proportion at round 0.30.
- This tells us that those who got a score of **B** refers to a loan that is considered to be of moderate risk, meaning the borrower has a higher likelihood of defaulting compared to a borrower with a higher credit grade (like A), but is still considered less risky than a "C" or "D" grade loan.
- And the next highest is **A** followed by **C.**
- The least is **G**; G loans have the highest expected risk of loss. Accordingly, G loans pay the highest interest rate in order to compensate lenders for the increased risk relative to an A loan.



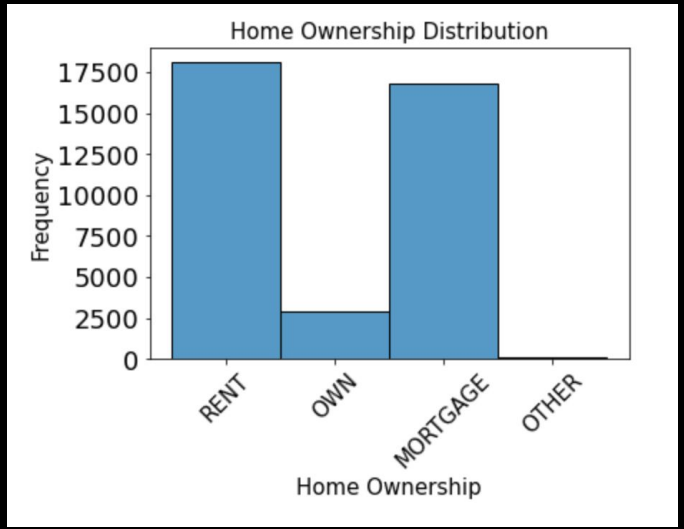
Inference:

- Employees who have an experience with 10 years are more present in the dataset.



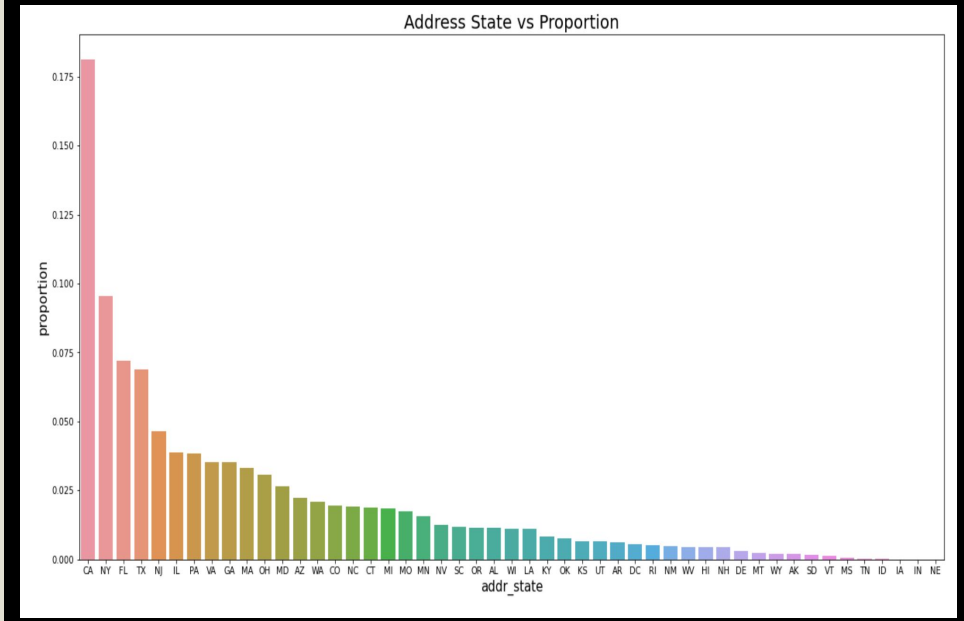
Inference:

- Based on the Home Ownership Distribution, we can see that customers take more **RENT** and **MORTGAGE** loans compared to **OWN** and **OTHER** with regards to home ownership loans.
- This can be due to variety of reasons, but generally those who **RENT** a house have a frequency of 17,500 in the dataset compared to those who take a loan for **MORTGAGE** having a frequency of 16,000.



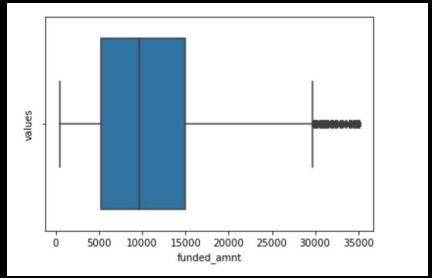
Inference:

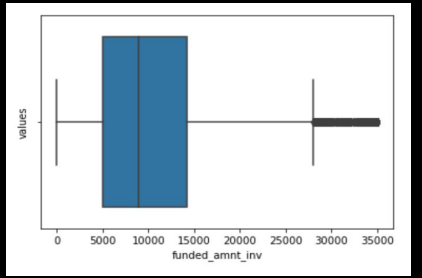
- Based on the Address State vs Proportion graph to the right, we can see that those who live in **CA**, **NY,** and **FL**.
- **CA:** 0.175
- **NY:** 0.980
- **FL:** 0.070
- Based on the this, we can infer that it is true as generally goods, commodities, houses, etc are more expensive near the coastal areas such as the states we have analyzed.

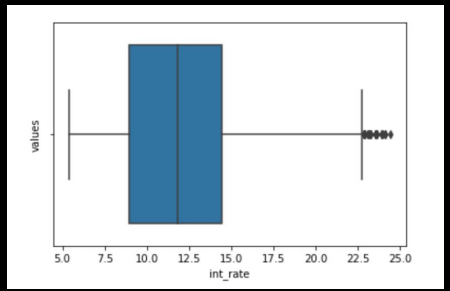


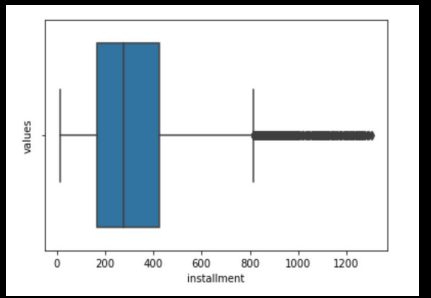
Inferences:

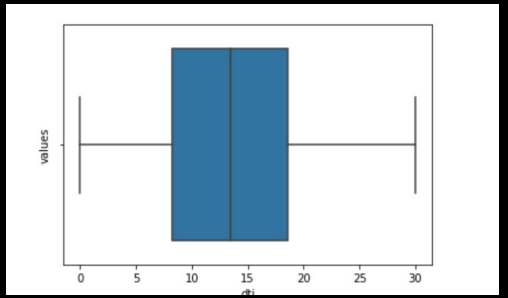
1. The annual_inc of most of the loan applicants is between 35,000 - 80,000
2. The loan amount of most of the loan applicants is between 5,000 - 15,000
3. The funded amount of most of the loan applicants is between 5,000 - 13,500
4. The funded amount by investor for most of the loan applicants is between 5,000 - 15,000
5. The interest rate on the loan is between 9% - 14.5%
6. The monthly installment amount on the loan is between 180 - 430
7. The debt to income ratio is between 8 - 18





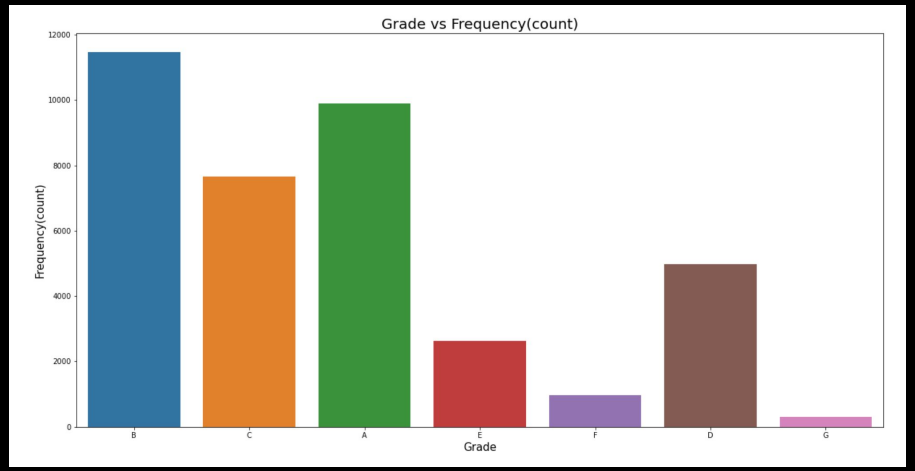






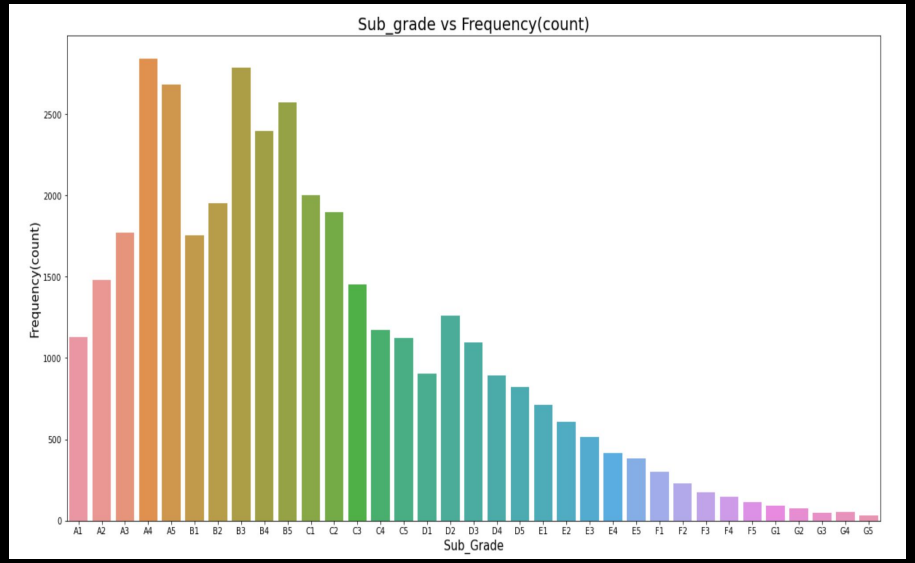
Inference:

- Based on the graph, “Grade vs Frequency(count)”, we can infer that those who got a grade **B** and **A** are more likely to receive loans it means:
  - **B:** A grade B loan indicates moderate credit risk with limited safety for ongoing payments. While the borrower meets current commitments, their ability to pay may be affected by economic or business changes.
  - **A:** loan signifies the lowest risk category, indicating the borrower is highly creditworthy and likely to repay on time. This grade offers the best loan terms, including lower interest rates and favorable conditions.



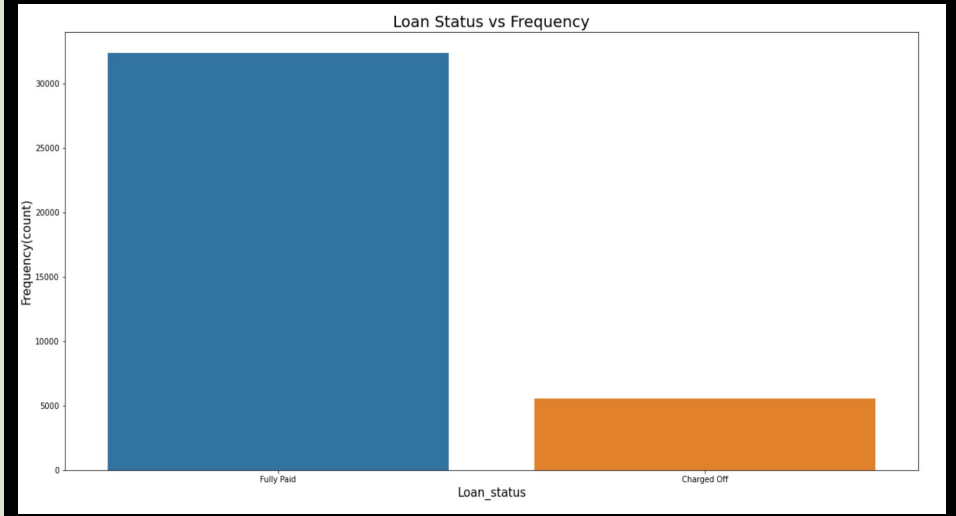
Inference:

- Sub_Grade **A1**, **B1**, **B3**, **B5** have the highest frequency count, which means that many loans are given on these 4 grades compared to the others.
- However if we look at the distribution, we can see that it is skewed to the left a bit and is not normally distributed which indicates those who got a high range grade got a higher loan compared to those with a low grade.



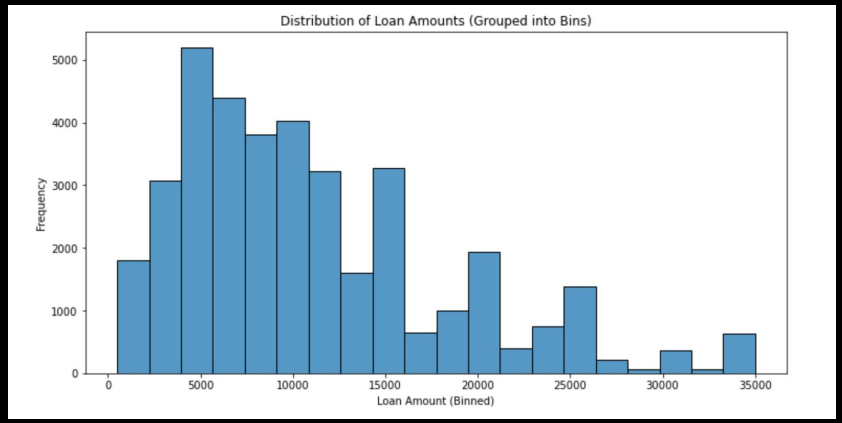
Inference:

- Based on the graph Loan Status vs Frequency, we can see that those who **Fully Paid** are higher (30,000+) compared to those who are **Charged off** (5000+).



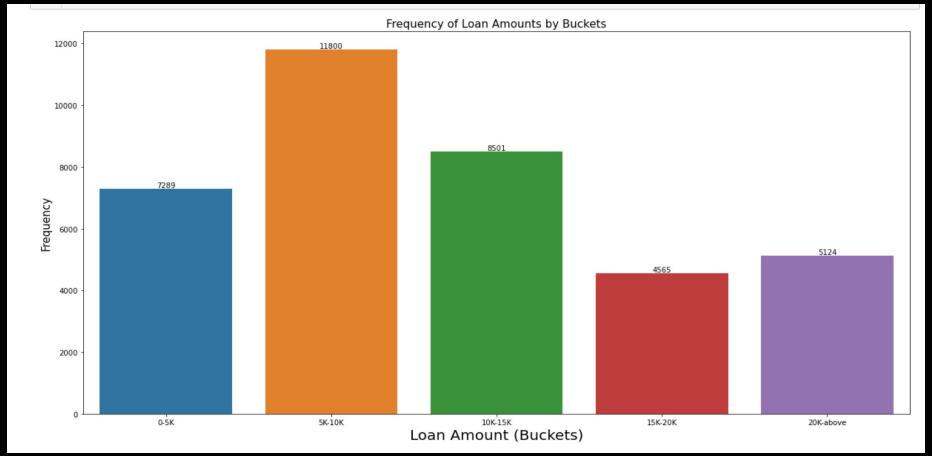
Inference:

- Based on the graph, Distribution of Loan Amounts (Grouped into Bins) we can infer that the frequency of those who took loan amount between 5,000-15,000 USD are much more compared to the rest of bins.
- However, this analysis is not a factor to consider when giving a loan or not, and understanding who will default.
- This analysis will tell the bank approximately what is the average loan amount range taken.



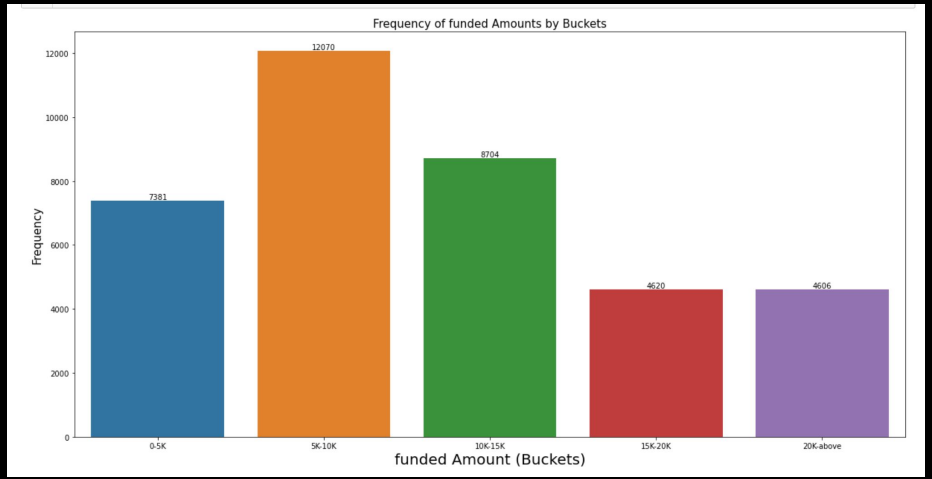
Inference:

- Based on the previous graph, we have plotted them into buckets and we can see that the loan amount for 5,000-10,000 USD range is higher with 11,800 customers.
- And the lowest accounts for 15,000-20,000 USD range with a frequency of 4565 customers.
- Now based on the previous graphs we can say that those who have a 10 year employment history are likely to take larger loans, thus consolidating our inference.



Inference:

- The Funded amount is highest in the range of 5,000-10,000 USD with 12,070 customers.
- And the lowest is 20,000+ USD with a frequency of 4606.



Univariate Analysis Insights

1. Certain purposes, like "credit card" and "small business," might be associated with a higher risk of default due to larger loan amounts and potentially higher risk profiles of borrowers.
2. Debt consolidation loans might be associated with a higher risk of default if borrowers are unable to manage their finances effectively after consolidation.
3. Employees who have an experience of 10+ years will more likely take higher loan amounts (approx 30,000-35,000 USD) compared to those who have less years of experience, i.e. below 3 years.
4. Applicants with higher grades (A and B) might pose a lower risk of default, as they have demonstrated consistent academic performance.
5. Applicants with lower grades (E, F, and G) might have a higher risk of default, as their academic performance suggests potential financial challenges or lower motivation.
6. California topped the list with 1,055 loan defaults, suggesting that the lending company should implement more stringent eligibility requirements or credit evaluations for applicants from this state to mitigate future risks.
7. The monthly installment amount on the loan is between 180 - 430
8. Applicants with higher sub-grades (A1, A2, etc.) are likely to have a lower default risk, as these sub-grades are associated with better creditworthiness and lower risk profiles.
9. Applicants with lower sub-grades (F1, F2, etc.) are more likely to have a higher default risk, as these sub-grades are associated with poorer creditworthiness and higher risk profiles.
10. The terms of the loan, such as interest rate, repayment period, and fees, can also influence default risk. Higher interest rates and shorter repayment periods can make it more difficult for borrowers to repay their loans.

5.Segmented Univariate Analysis

Segmented univariate analysis involves analyzing a single variable **within the context of one or more segments or groups** defined by another variable. The goal is to observe how the distribution, central tendency, or variability of the primary variable changes across different segments.

Key Features of Segmented Univariate Analysis

1. **Primary Variable:** Focuses on one variable (numeric or categorical).
2. **Segmentation Variable:** The variable used to divide the data into meaningful groups (e.g., loan status, gender, grade).
3. **Purpose:** Understand patterns, trends, or variations within subsets of the data.

Why Segment a Univariate Analysis?

- To **compare distributions** or trends across groups.
- To identify **group-specific behaviors or anomalies**.
- To provide deeper insights that may not be evident in an overall analysis.

Benefits of Segmented Univariate Analysis

1. **Granular Insights:** Uncover patterns that are unique to specific groups.
2. **Better Decision-Making:** Helps target interventions for specific segments.
3. **Exploratory Power:** A stepping stone for identifying relationships for bivariate/multivariate analysis.

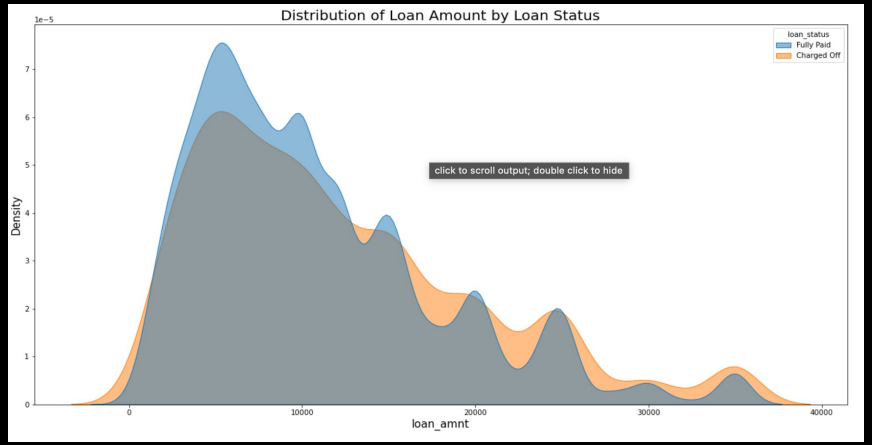
Inferences:

Fully Paid Loans (Blue):

- The density is highest for loans with smaller amounts, peaking around $10,000. This suggests that borrowers with lower loan amounts are more likely to repay their loans fully.
- As the loan amount increases beyond $20,000, the density gradually decreases, indicating that higher loan amounts are less frequently fully repaid.

Charged-Off Loans (Orange):

- Charged-off loans show a significant overlap with fully paid loans in the lower loan amount range ($5,000–$15,000).
- However, the density of defaults rises relative to fully paid loans as loan amounts exceed $15,000. This highlights that larger loans are riskier.



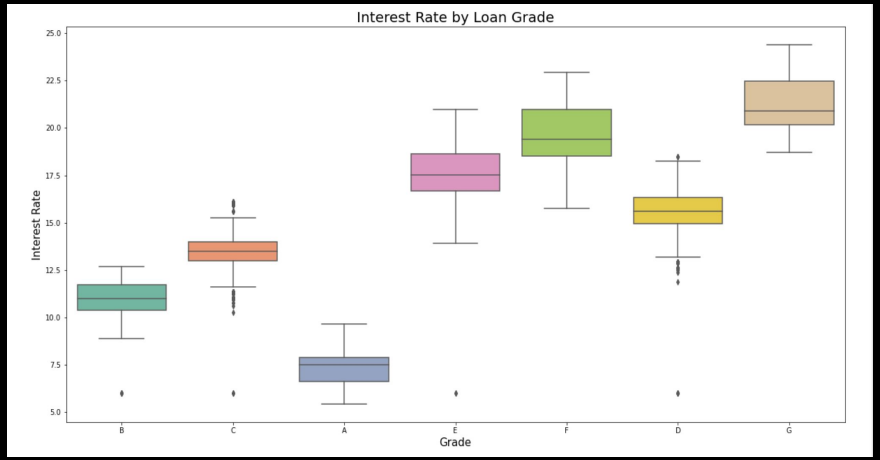
Inferences:

Trends Across Grades:

- Interest rates increase steadily from Grade A to Grade G, reflecting the increasing credit risk associated with lower grades.
- Grade A loans have the lowest median interest rate (~6%), while Grade G loans exhibit the highest median rate (~22%).

Spread and Outliers:

- Grades B, C, and D show moderate variability in interest rates, suggesting diverse borrower profiles within these grades.
- Grades A and G have tighter spreads, indicating consistent risk profiles among the borrowers in these categories.
- Outliers are present across all grades, particularly in Grades B and F, which might represent borrowers with unusual credit circumstances.



Inferences:

Highest Loan Amounts:

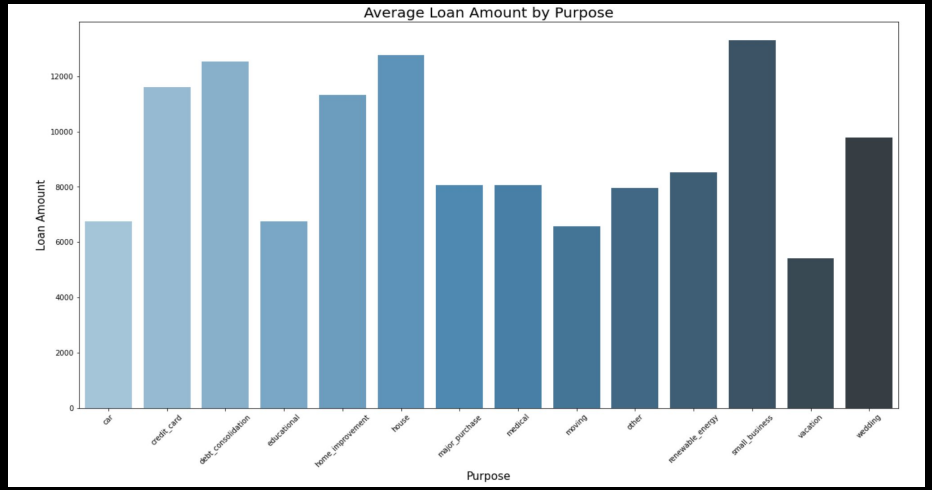
- Borrowers for **"small business"** and **"house"** purposes take the highest average loan amounts, exceeding $12,000.
- These categories likely indicate significant financial commitments, and the associated risks could depend on external factors like market stability or personal income.

Moderate Loan Amounts:

- Borrowers in these categories may also have moderately higher risk, as consolidating debts or funding large purchases often occurs under financial strain.

Lowest Loan Amounts:

- These purposes might involve more discretionary spending, suggesting lower risk of default due to the smaller loan sizes.



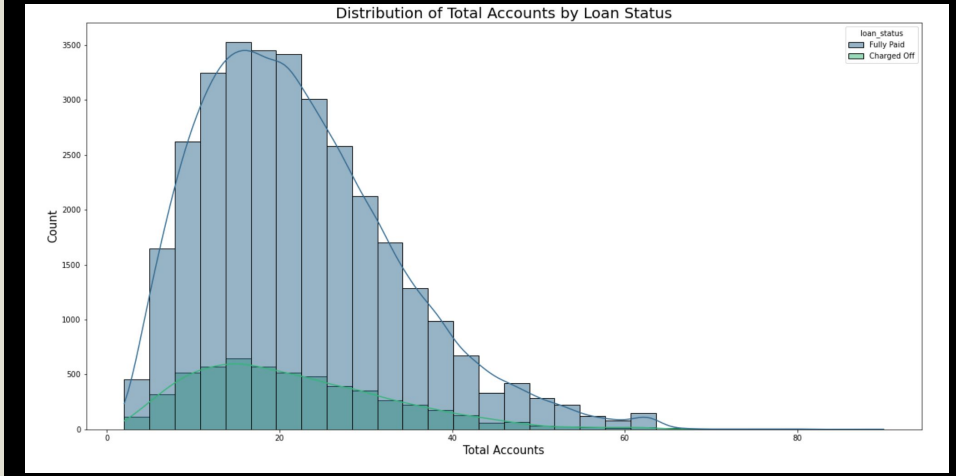
Inferences:

Fully Paid Loans:

- Most fully paid loans are concentrated in borrowers with **10–40 total accounts**.
- This indicates that borrowers with moderate credit experience are more likely to successfully repay loans.

Charged-Off Loans:

- Charged-off loans are more frequent among borrowers with **fewer total accounts** (0–20 accounts).
- Borrowers with limited credit experience may struggle with managing financial obligations, increasing the likelihood of default.

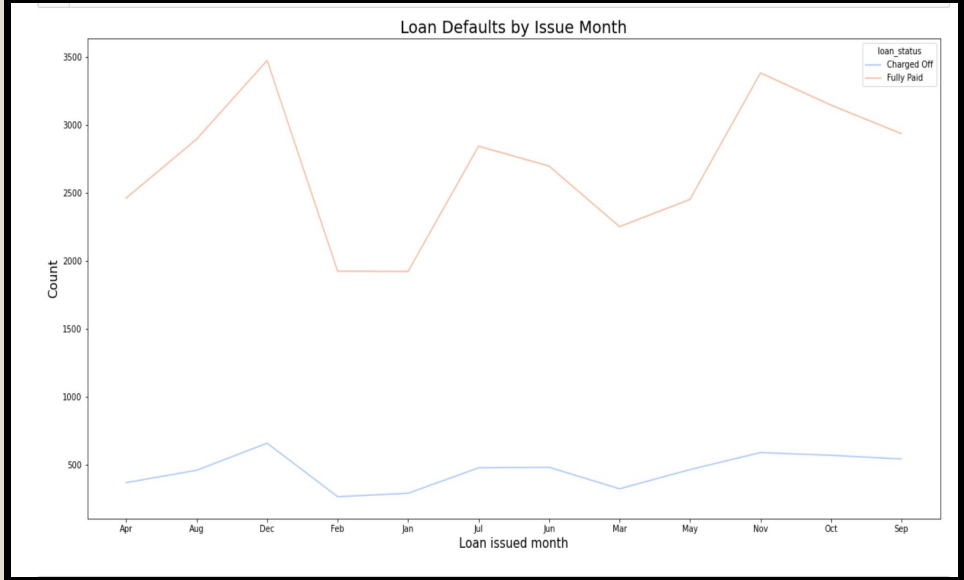


Inference:

- Based on the graph, “Loan Defaults by Issue Month” we can infer that:

1. **Fully Paid:**
   a. Sudden spike to pay towards the end of the year.
   b. Declining in the start of year (Jan, Feb, March) as loans are issued at the start of every financial year.
   c. No change in the middle months (April,May,June,July)

2. **Charged Off:**
   a. Customers are more likely to default towards the end of the year as payments towards the end of the year need to be paid and thus an increasing interest rate.
   b. Frequency of customers will decrease in the beginning of the year



Segmented Univariate Analysis Insights

1. A majority of loans are concentrated around the $10,000 mark, suggesting that smaller loan amounts are more likely to be fully repaid.
2. Loans in the $5,000–$15,000 range exhibit a significant overlap between those that were fully repaid and those that were charged off.
3. Grade A loans have the lowest median interest rate (~6%), while Grade G loans exhibit the highest median rate (~22%).
4. Individuals with a smaller number of credit accounts (0-20) are more likely to default on their loans, suggesting that limited credit history can contribute to financial instability.
5. Borrowers seeking loans for "small business" or "house" purposes typically take on the highest average amounts, exceeding $12,000. These categories reflect substantial financial commitments, with associated risks potentially influenced by factors such as market stability and personal income.
6. Towards the end of the year, customers may face heightened financial pressures from holiday spending and potential interest rate increases, which could lead to a higher risk of default.

6.Bivariate Analysis

Bivariate analysis examines the relationship between two variables to identify patterns, associations, or dependencies. It helps answer questions about how one variable changes in relation to another.

Key Features of Bivariate Analysis

1. **Two Variables:** Focuses on the interaction or relationship between two variables.
2. **Types of Variables:**
    - **Numeric-Numeric:** Both variables are continuous.
    - **Categorical-Categorical:** Both variables are discrete categories.
    - **Numeric-Categorical:** One variable is continuous, and the other is categorical.
3. **Purpose:** To understand correlations, dependencies, or trends between the variables.

Why is Bivariate Analysis Important?

1. **Understanding Relationships:** Helps determine whether variables are related and in what way.
2. **Exploratory Insights:** Identifies key relationships that may inform predictive modeling or business decisions.
3. **Guides Hypotheses:** Provides evidence to support or refute hypotheses about variable interactions.

Inferences:

High-Grade Loans (A, B):

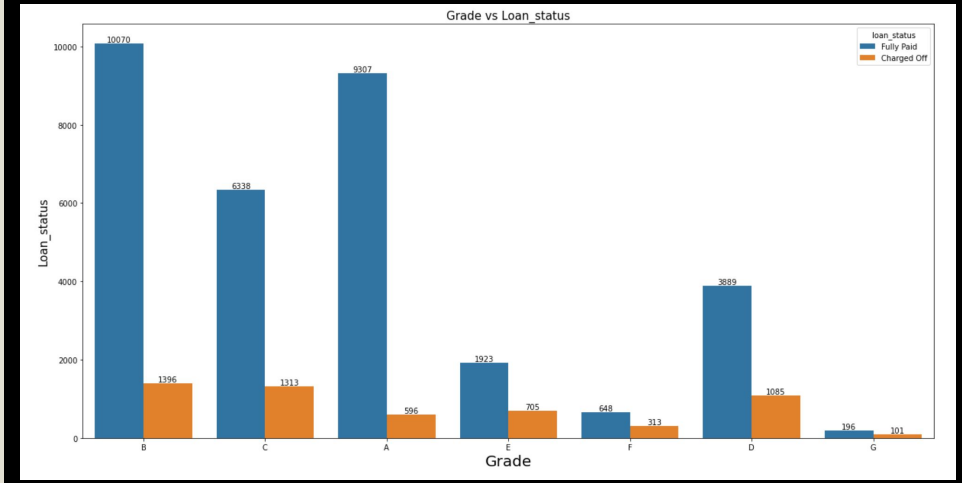
- Grades A and B have the largest number of fully paid loans, suggesting that borrowers in these grades are the least risky.
- Default rates (charged off) are low relative to the total loans issued in these grades.

Mid-Grade Loans (C, D):

- Grades C and D show a noticeable increase in defaults compared to A and B.
- This suggests that mid-tier grades involve higher risk, likely reflecting moderate creditworthiness.

Low-Grade Loans (E, F, G):

- Grades E, F, and G exhibit significantly higher default rates relative to the number of loans issued.
- These grades are associated with high-risk borrowers, indicating the need for stricter monitoring and higher interest rates.



Inferences:

High-Sub-Grades (A1, A2, A3, etc.):

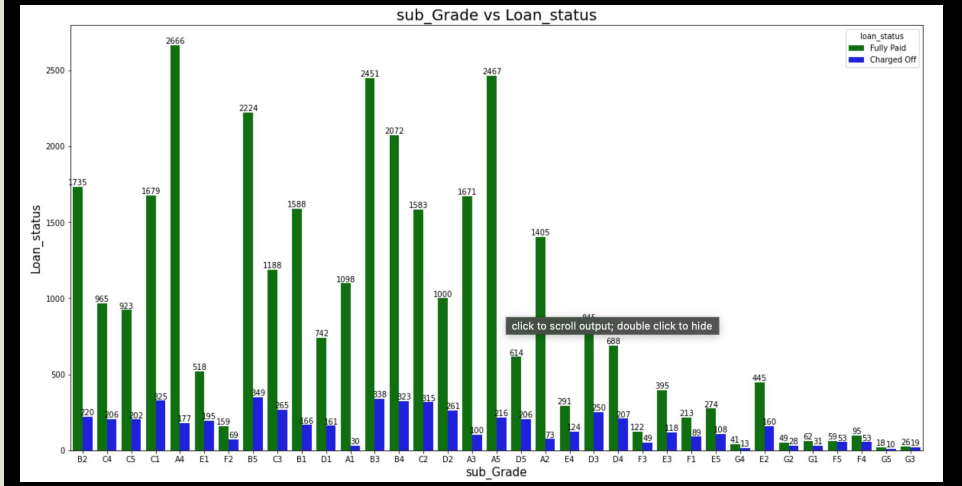
- Sub-grades in the A tier (e.g., A1, A2) show the highest proportion of fully paid loans with very low defaults, emphasizing that borrowers in these sub-grades are highly reliable.

Mid-Sub-Grades (B3 to C5):

- Default rates increase gradually within sub-grades as the risk level rises. For instance, B4 and C4 show a notable rise in defaults compared to A-tier sub-grades.
- This progression highlights the granular nature of risk assessment.

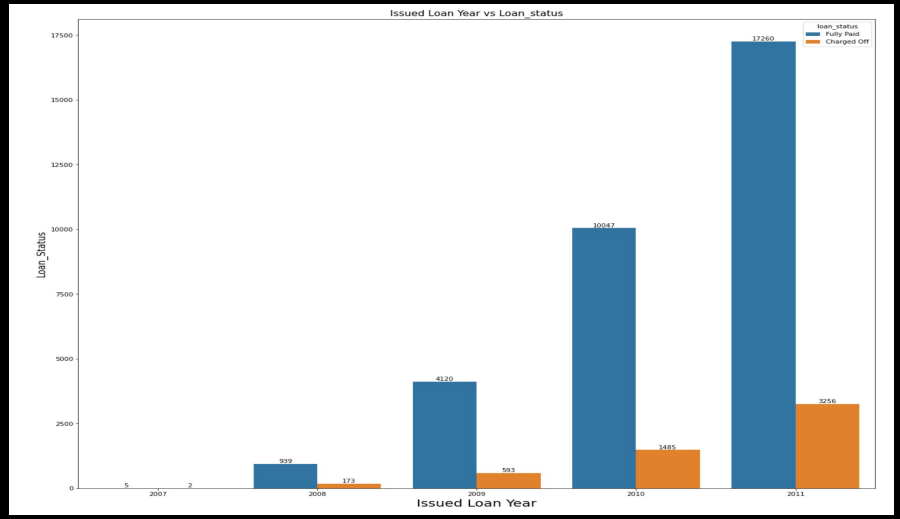
Low-Sub-Grades (E3, F5, G3, etc.):

- Sub-grades in the F and G categories exhibit very high default rates, with fewer fully paid loans.
- These loans are high-risk and likely require additional guarantees or stricter underwriting processes.



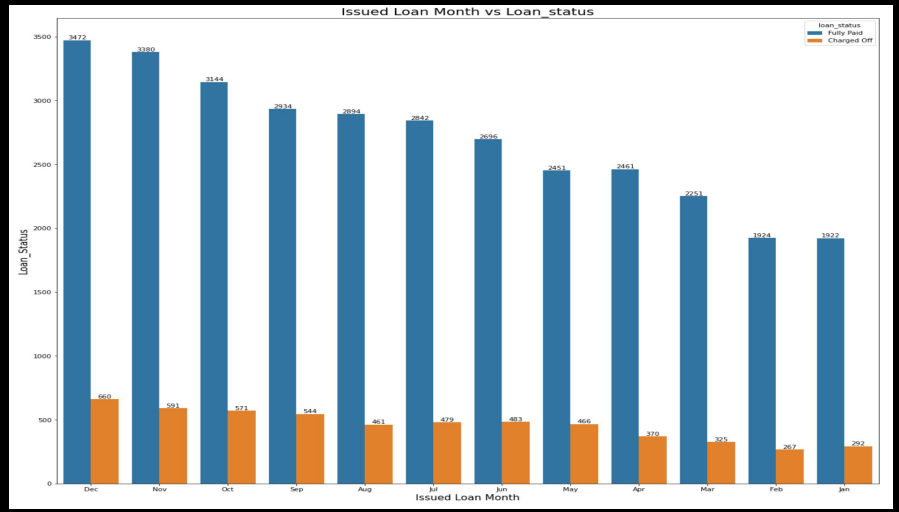
Inferences:

- Those who got issued a loan in 2011 had more fully paid customers compared to those who defaulted.
- As the year is increasing, the number of fully paid customers are increasing more than the customers who defaulted which indicates a positive growth.



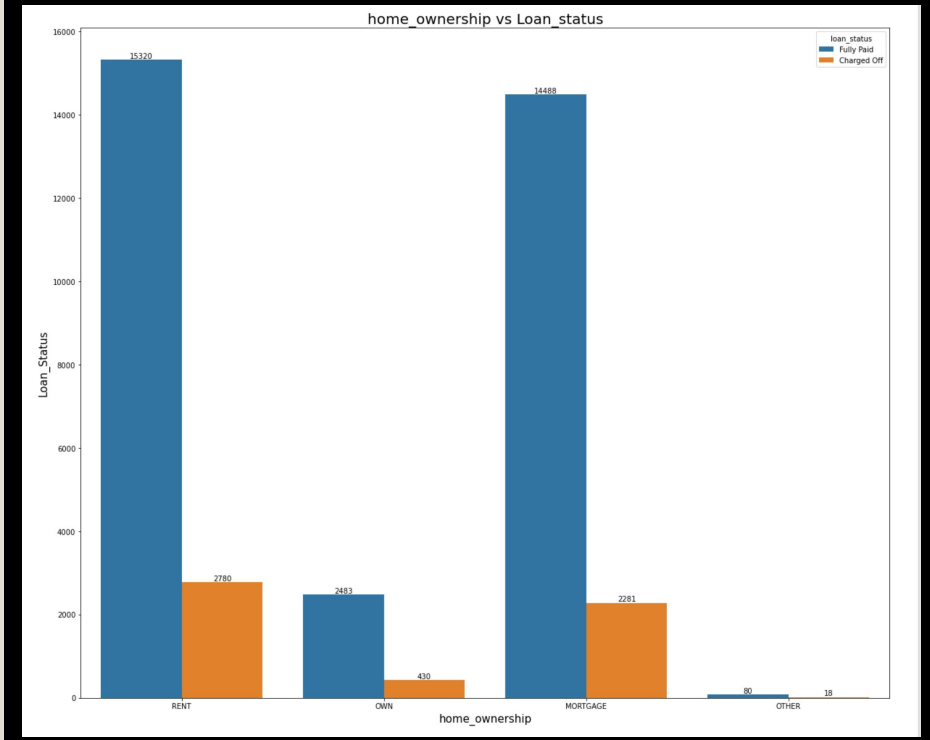
Inferences:

- According to the graph, those who took a loan in december are more likely to pay to it off compared to those who took a loan at the beginning of the year.
- This is true because towards the end of the year many customers will pay back.
- On the other hand, those who defaulted is the most in december because they are unable to pay the installments or due to a higher interest rate, they are unable to pay it off.



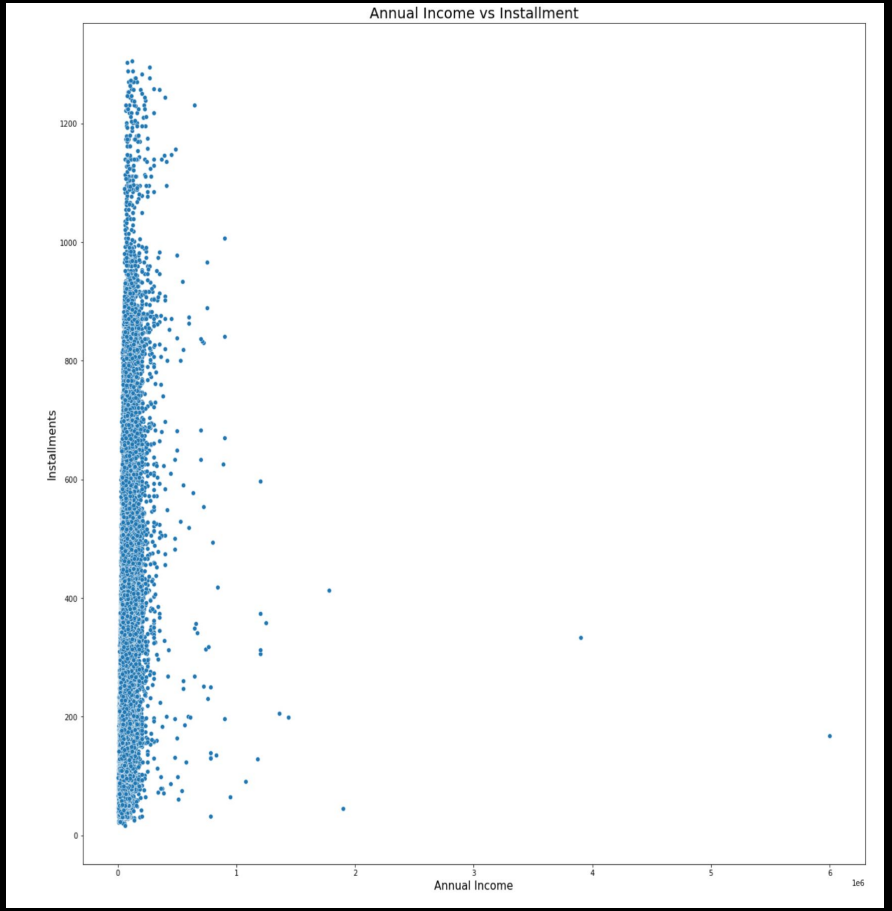
Inferences:

- Based on the graph, we can infer that those who a **RENT** a home tend to have a higher loan status of 15,320 and fully pay it off compared to those who charged off have a loan status of 2,780.
- The next highest is **MORTGAGE**, and this shows that those who fully paid have a higher loan status of 14,448 and charged off is 2,281.
- And those who **OWN** their own home have the least loan status as they do not need to take a loan since they are the property owners.



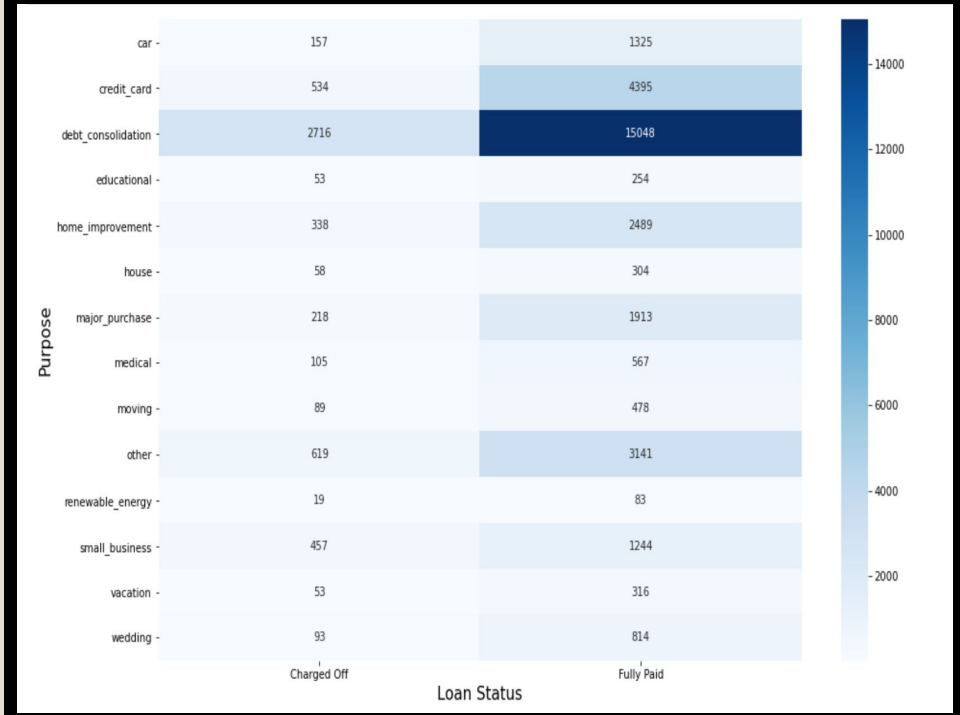
Inferences:

- **Annual Income and Installments:** The scatter plot shows a general trend where higher annual incomes are associated with larger loan installments. However, there is also a significant amount of variation, with some individuals having high incomes but low installments, and vice versa.
- **Outliers:** There are a few data points with very high annual incomes but relatively low installments. These could be potential outliers or might represent specific scenarios like early loan repayment or special loan terms.
- **Income-to-Installment Ratio:** The relationship between annual income and installment amount suggests that a higher income-to-installment ratio might be associated with a lower risk of default. However, this needs to be analyzed in conjunction with other factors like debt-to-income ratio, credit history, and loan purpose.



Inferences:

1. **Debt Consolidation:** This category has the highest number of both Charged Off and Fully Paid loans, indicating a significant portion of the loan portfolio
2. **Credit Card:** This category also has a substantial number of loans, with a noticeable difference between Charged Off and Fully Paid counts, suggesting potential risk factors
3. **Other:** This category, likely encompassing various miscellaneous purposes, shows a relatively high number of Charged Off loans compared to Fully Paid
4. **Small Business:** While having a significant number of loans, the ratio of Charged Off to Fully Paid seems relatively balanced.
5. **Home Improvement:** This category shows a lower number of loans compared to some others, but the ratio of Charged Off to Fully Paid is relatively high.
6. **Medical:** This category, while not having a large number of loans, shows a higher proportion of Charged Off loans compared to Fully Paid.



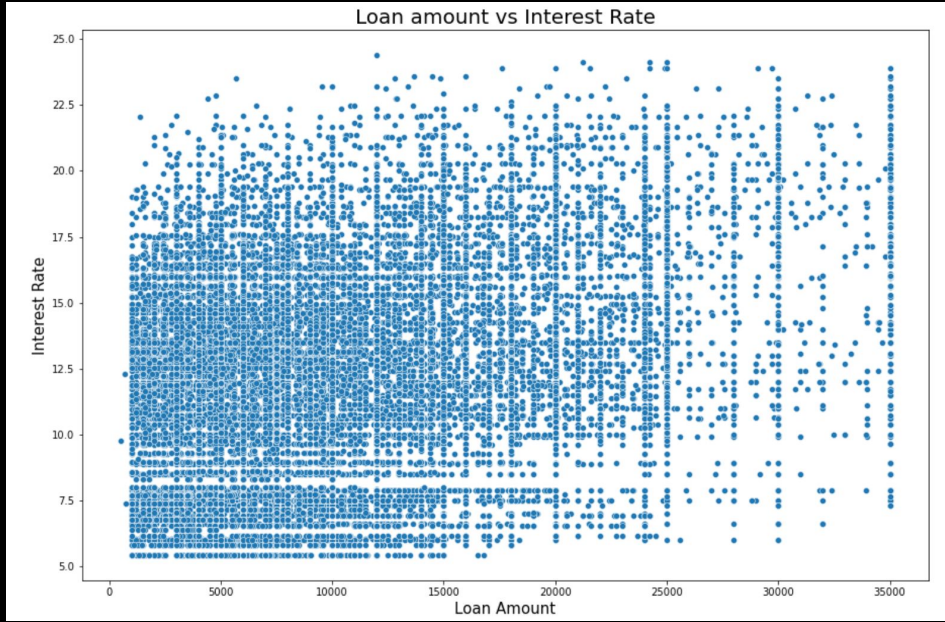
Inferences:

1. **Inverse Relationship**: Higher loan amounts generally have lower interest rates, indicating better creditworthiness for larger loans.

2. **Dense Cluster**: Most loans range between $5,000–$15,000 with interest rates of 10–20%.

3. **High Risk Segment**: Small loans (<$10,000) often come with high interest rates (>15%), reflecting higher risk.

4. **Low Risk Segment**: Large loans (>$20,000) tend to have lower interest rates (<10%), showing lower-risk borrowers.



Inferences:

Loan Amount and Annual Income:

- There is a moderate positive correlation (0.27) between loan_amnt (loan amount) and annual_inc (annual income). This suggests that individuals with higher incomes tend to take larger loans.

Loan Amount and Interest Rate:

- A moderate correlation (0.3) exists between loan_amnt and int_rate (interest rate). This indicates that higher loan amounts are slightly associated with higher interest rates.

Debt-to-Income Ratio (DTI) and Annual Income:

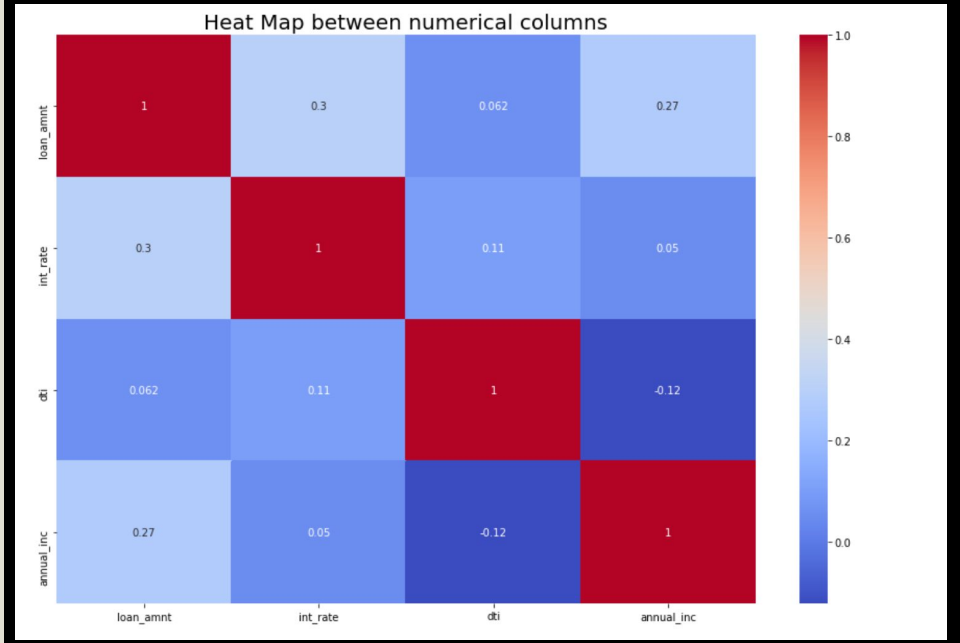
- A weak negative correlation (-0.12) between dti and annual_inc indicates that higher incomes are slightly associated with lower debt-to-income ratios.

Debt-to-Income Ratio (DTI) and Loan Amount:

- A very weak positive correlation (0.062) suggests little to no direct relationship between the size of the loan and the debt-to-income ratio.

Interest Rate and Other Variables:

- The interest rate (int_rate) shows weak correlations with other variables (loan amount: 0.3, annual income: 0.05, DTI: 0.11). This implies that the factors considered here only slightly influence interest rates.

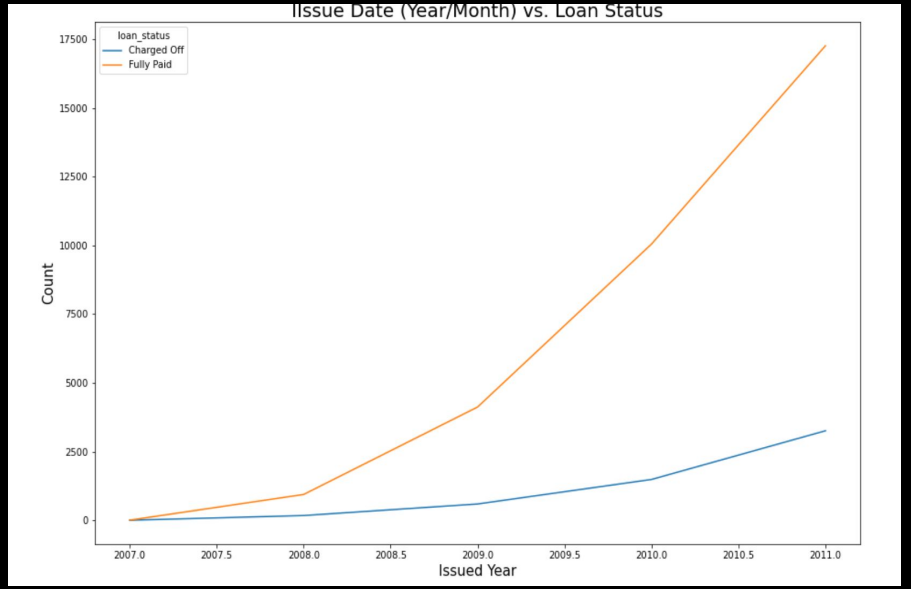


Inferences:

- **Increase in Loan Volume:** The plot shows a clear upward trend in the number of loans issued over time, with a significant increase from 2007 to 2010.
- **Charged Off vs. Fully Paid:** The number of both Charged Off and Fully Paid loans has been increasing over time, but the rate of increase for Charged Off loans appears to be slightly higher than for Fully Paid loans.
- **Potential Uptick in Default Rates:** The increasing gap between Charged Off and Fully Paid loans suggests that the default rate might be rising over time.

Potential Insights for Loan Default:

- **Economic Conditions:** The increase in loan volume and default rates might be influenced by macroeconomic factors like interest rates, unemployment, and inflation.
- **Lending Standards:** Changes in lending standards or underwriting criteria might have contributed to the increase in default rates.
- **Borrower Behavior:** Shifts in borrower behavior, such as increased risk-taking or reduced financial discipline, could be contributing to the rising default rates.



Bivariate Analysis Insights

1. High DTI ratios could be a strong indicator of potential default. Borrowers with a significant portion of their income already committed to debt might struggle to meet additional loan obligations.
2. Higher interest rates could increase the likelihood of default, especially for borrowers with lower incomes or higher DTI ratios.
3. While a large loan amount may not be a direct indicator of default, it could be a risk factor when combined with other factors like DTI and interest rate.
4. Grades A and B have the largest number of fully paid loans, suggesting that borrowers in these grades are the least risky.
5. Grades C and D show a noticeable increase in defaults compared to A and B.
6. This suggests that mid-tier grades involve higher risk, likely reflecting moderate creditworthiness.
7. Sub-grades in the A tier (e.g., A1, A2) show the highest proportion of fully paid loans with very low defaults, emphasizing that borrowers in these sub-grades are highly reliable.
8. The data indicates that loans issued in December tend to have a better repayment performance than those issued earlier in the year. This trend might be attributed to factors such as year-end bonuses or timely repayment to avoid higher interest charges. Conversely, December also witnesses the highest number of defaults, which could be due to increased financial burdens during the holiday season or difficulties in meeting increased interest payments.
9. Loans for debt consolidation, credit card, and other purposes might be associated with a higher risk of default.
10. those who a **RENT** a home tend to have a higher loan status of 15,320 and fully pay it off compared to those who charged off have a loan status of 2,780.
11. Small loans (<$10,000) often come with high interest rates (>15%), reflecting higher risk.

7.Multivariate Analysis

Multivariate analysis is a statistical technique used to examine and understand the relationships among **three or more variables** simultaneously. It helps identify patterns, correlations, and interactions within complex datasets and is often applied to explore how multiple variables jointly influence outcomes.

Key Characteristics

- **Multiple Variables:** Focuses on analyzing more than two variables at a time.
- **Interdependence:** Explores the relationships, interactions, or dependencies among variables.
- **Complexity:** Provides insights that bivariate or univariate analyses cannot capture.

Purpose of Multivariate Analysis

1. **Identify Relationships:** Understand how variables interact and influence each other.
2. **Predict Outcomes:** Develop models to predict a dependent variable using multiple predictors.
3. **Dimensionality Reduction:** Simplify datasets with many variables into fewer meaningful dimensions while retaining essential information.
4. **Uncover Patterns:** Detect trends, clusters, or groupings within the data.

Importance

Multivariate analysis is essential for exploring real-world problems where multiple variables interact, allowing for more accurate insights, decision-making, and predictions.

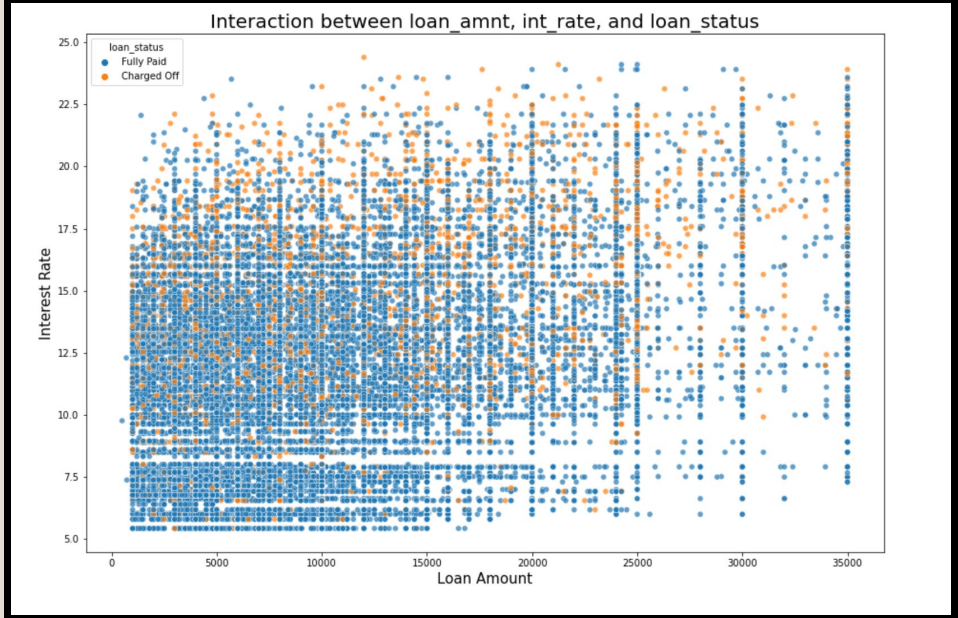
Inferences:

Interaction between Loan Amount, Interest Rate, and Loan Status:

- **Distribution:** The scatter plot shows a wide range of loan amounts and interest rates. The majority of loans are concentrated in the lower range of both loan amounts and interest rates.

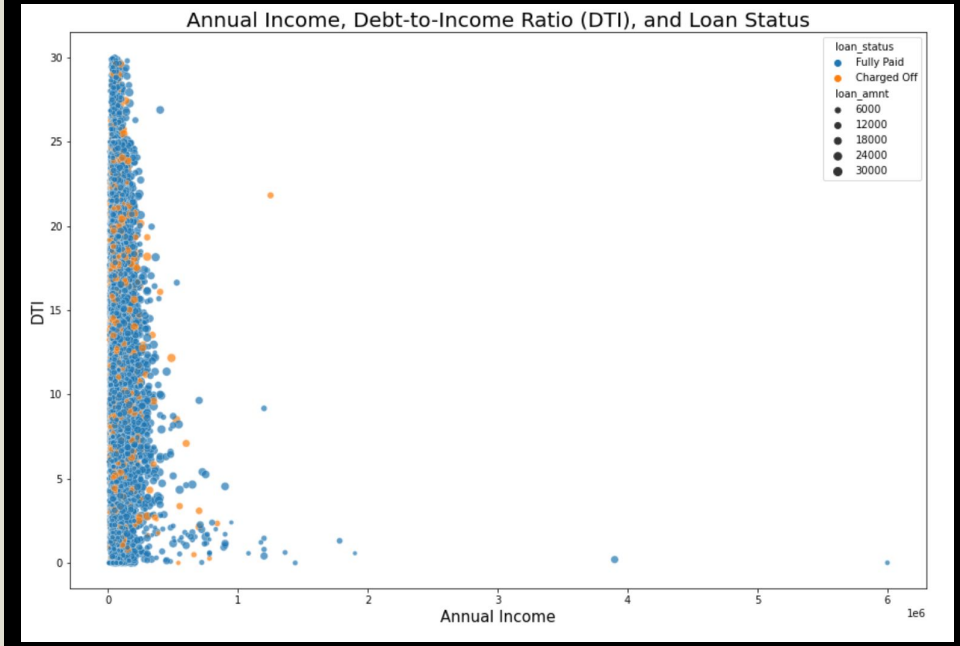
- **Default Risk:**

    - There seems to be a concentration of Charged Off loans in the region where loan amounts are higher and interest rates are higher. This suggests that loans with larger amounts and higher interest rates might be associated with a higher risk of default.
    - However, there are also some Charged Off loans scattered throughout the plot, indicating that default risk can be influenced by other factors besides loan amount and interest rate.



Inferences:

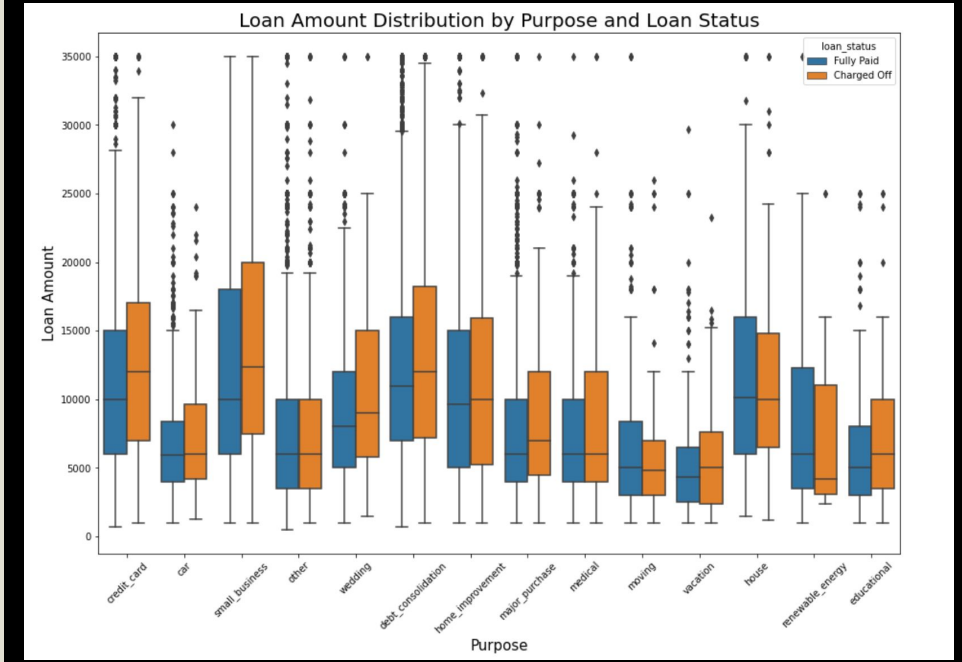
- There seems to be a wider range of annual incomes for both fully paid and charged-off loans, with no clear separation based on income alone.
- There's a noticeable cluster of charged-off loans in the higher DTI range. This suggests that a higher debt-to-income ratio might be a significant factor in loan defaults.



Inferences:

- The box plots show that loan amounts vary significantly across different purposes. Some purposes, like "credit card" and "small business," tend to have larger loan amounts compared to others like "wedding" and "educational."
- The box plots reveal that for most purposes, the median loan amount for Charged Off loans is higher than that for Fully Paid loans. This suggests that larger loan amounts might be associated with a higher risk of default, regardless of the purpose.

- The presence of outliers (individual data points outside the whiskers) indicates that there are some loans with significantly larger amounts compared to the majority within each purpose and loan status. These outliers might represent special cases or might be due to data errors.



Inferences:

Loan Amount:

- **Distribution:** The distribution of loan amounts is right-skewed, with a majority of loans being smaller amounts.
- **Default Risk:** Larger loan amounts seem to be associated with a slightly higher proportion of Charged Off loans, suggesting a potential link between loan size and default risk.

Interest Rate:

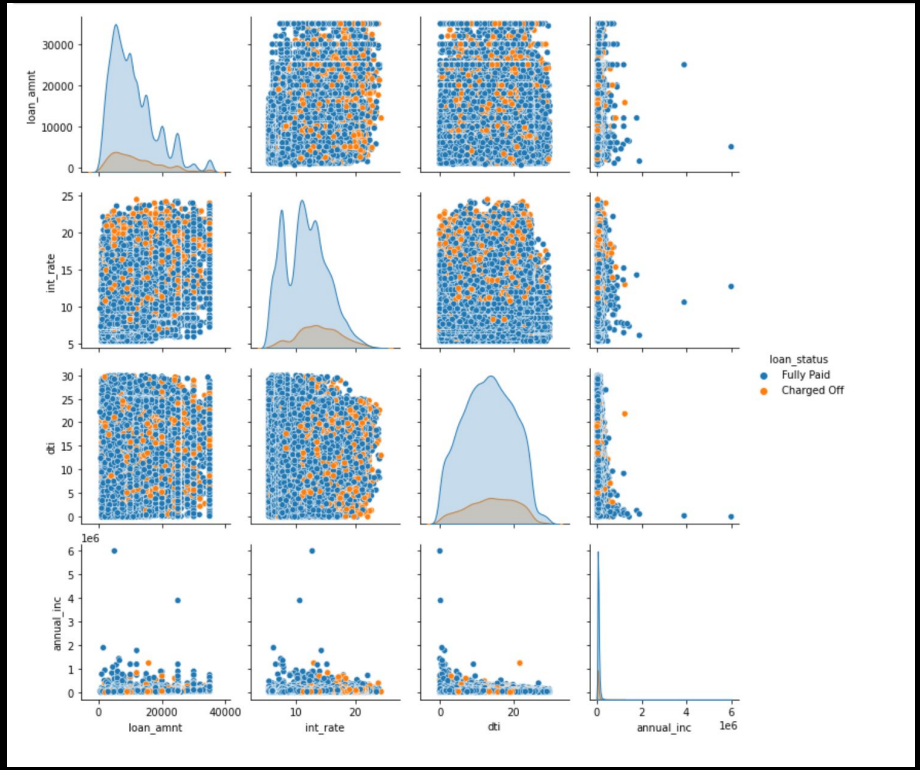
- **Distribution:** Interest rates are concentrated in the lower range, with a few outliers on the higher end.
- **Default Risk:** Higher interest rates are associated with a significantly higher proportion of Charged Off loans, indicating a strong link between interest rate and default risk.

Debt-to-Income Ratio (dti):

- **Distribution:** dti is also right-skewed, with most borrowers having a relatively low dti.
- **Default Risk:** Higher dti is associated with a higher proportion of Charged Off loans, suggesting that borrowers with higher debt burdens might be more likely to default.

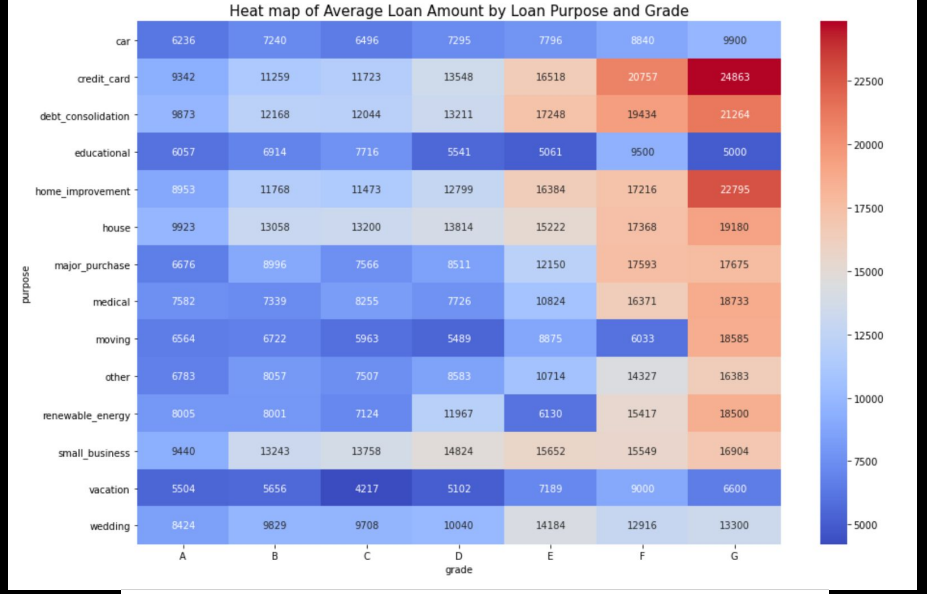
Annual Income:

- **Distribution:** Annual income is also right-skewed, with a majority of borrowers having lower incomes.
- **Default Risk:** There doesn't seem to be a clear relationship between annual income and default risk. However, the distribution of Charged Off loans across income levels might warrant further investigation



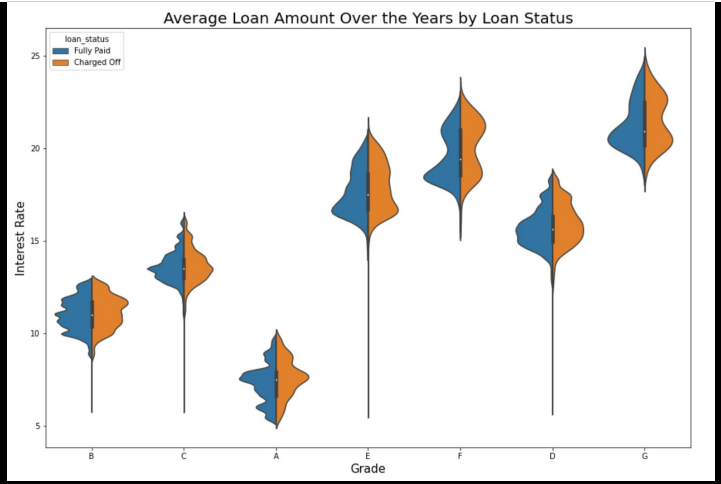
Inferences:

- **Increase in Loan Amount Over Time:** The heatmap shows a general trend of increasing average loan amounts across all grades and loan statuses over the years. This suggests that lenders might have been approving larger loans over time.
- **Higher Loan Amounts for Charged Off Loans:** Within each grade, the average loan amount for Charged Off loans is consistently higher than for Fully Paid loans. This indicates that larger loans might be associated with a higher risk of default.
- **Grade-wise Variation:** The average loan amounts vary significantly across different grades. Lower grades (A, B) tend to have lower average loan amounts, while higher grades (F, G) have significantly higher average loan amounts.
- The increasing average loan amounts over time might be a contributing factor to the potential rise in default rates.



Inferences:

- **Interest Rate and Default Risk:** The graph suggests that higher interest rates might be associated with a higher risk of default, regardless of the borrower's creditworthiness (grade).
- **Grade as a Risk Indicator:** Lower grades (A, B) appear to be associated with lower default rates, while higher grades (F, G) might be riskier. However, the absolute risk of default is still higher for loans with higher interest rates within lower grades.



Multivariate Analysis Insights

1. The data suggests a correlation between loan size and interest rate with the likelihood of default. Larger loans with higher interest rates seem to be more prone to charge-offs.
2. A high DTI ratio is a strong indicator of potential default. Borrowers with a significant portion of their income already committed to debt might struggle to meet additional loan obligations.
3. Larger loan amounts, especially when combined with a high DTI, could increase the risk of default.
4. Larger loan amounts, regardless of purpose, appear to be associated with a higher risk of default, example is those who take a loan of 20,000+ USD tend to default more than those who took a loan between the range of 10,000-15,000 USD
5. Interest rates are concentrated in the lower range, with a few outliers on the higher end.
6. As the loan grade increases from A to G, the average loan amount generally increases. This suggests that borrowers with lower credit scores tend to receive smaller loans, likely due to higher perceived risk.
7. For **Debt Consolidation** and **Credit Card** loans, even the lower grades (D, E, F, G) have relatively high average loan amounts compared to other loan purposes. This suggests that these categories might be riskier, as borrowers with lower credit scores are still able to obtain larger loans.
8. **Educational** and **Moving** loans have the lowest average loan amounts, suggesting that these loans are typically smaller in size.

8.Conclusion

Suggestions:

1. Implement a robust risk assessment framework that considers multiple factors, including income, debt-to-income ratio, credit history, and loan purpose, to identify high-risk loans.
2. Overall, the univariate analysis suggests a positive trend in the academic performance of the applicant pool. However, a more comprehensive analysis, incorporating other relevant factors, is necessary to make informed decisions about loan approval and risk assessment.
3. To gain a deeper understanding of the relationship between loan amount, funded amount, and default risk, it would be beneficial to analyze the data further by considering factors like:
    - Borrower demographics (age, income, occupation)
    - Loan purpose
    - Creditworthiness metrics (credit score, debt-to-income ratio)
    - Economic indicators (interest rates, unemployment rates)
4. Anticipate peak periods where default will most likely occur based on the analysis and prevent it so the company does not incur a loss
5. Monitor and adjust interest rates
6. Consider income levels
7. Evaluate for those who took a loan for 60 months, as they are the customers who tend to default more.

THANK YOU